Supplementary Information Guide

SupplementaryTables.xlsx

Supplementary Table 1 | Placed APG contigs. Contigs are reported with their length, placement location, orientation, and the number of individuals containing the contig. If only one end of the contig was placed, a "." appears in either the start or end position to indicate the position is unknown. Additionally, if an alignment to GRCh38.p10 with at least 80% identity and 50% contig coverage was found, the alignment scaffold and position (using RefSeq identifiers for GRCh38 scaffolds), alignment identity, and contig coverage are provided.

Supplementary Table 2 | Unplaced APG contigs. Unplaced APG contig names, lengths, and number of individuals containing the contig are reported. If an alignment to GRCh38.p10 with at least 80% identity and 50% contig coverage was found, the alignment scaffold and position (using RefSeq identifiers for GRCh38 scaffolds), alignment identity, and contig coverage are provided.

Supplementary Table 3 | Putative APG Placements. Putative linking locations for unplaced contigs are reported if greater than 50% of linking mates with a mapping quality > 10 for a given end pointed to the same region, where a region was defined by grouping mates within 2 kb of each other. The total number of high quality mates, the percentage of those mates linking to that location, and the location are reported for each end of the contig.

Supplementary Table 4 | Gene intersections. Placed contigs intersecting an exon, mRNA, or lncRNA annotation are listed along with the annotation type and gene name. If a contig intersects multiple features, the contig is listed once per feature it intersects.

SupplementaryDataSet1.csv

Contains a matrix of 125,715 rows (contigs) by 910 columns (CAAPA individuals) with a 0 or 1 entry labeling whether the individual contained the contig. Row headers are the APG contig names. Column headers are the Illumina identifiers used in the dbGap deposition of the CAAPA genomes for each individual.

Supplementary Note 1

PopIns placements. Of the 1,246 one-ended placements, our pipeline found 1,229, while PopIns [16] found an additional 17 and confirmed 60 (see Methods). Of the 302 pan-genome contigs for which both ends were placed, 70 were placed by both our method and PopIns, 129 by our method only, and 103 by PopIns only. Those placed solely by PopIns were verified via alignment of contig ends.

PopIns was able to resolve placement locations for some insertions where our method, which uses both contig and mate-pair alignments, gave ambiguous results. This is likely a result of PopIns' utilization of split-read alignments to determine exact placement locations, which provides some additional power. However, our approach has advantages when the contigs shared between individuals are more divergent, as they tend to be in the CAAPA populations. PopIns merges similar sequences prior to attempting placement but excludes from further analysis contigs which have a partial, but insufficient (for merging) match to many other contigs. This resulted in PopIns considering only 5% of the full set of assembled contigs (81,650 of 1,536,049), while our approach placed many of these “unmergeable” contigs by placing all contigs first, and then clustering based on location. While this pre-merging step prior to placement worked well for the highly homogenous Icelandic population for which PopIns was designed, it was less effective for our more heterogenous African-descended populations.

Supplementary Note 2

Presence of African pan-genome variants in other populations. To examine the presence/absence of our APG insertions in the SGDP individuals, we performed analysis on 12 individuals from 6 European populations and 12 individuals from 6 continental African populations from the Simons Genome Diversity Project (SGDP)[27]. After assembling the unaligned reads from these 24 genomes, we compared the assembled contigs to the APG sequences to determine how many of the APG sequences were present in the SGDP genomes. The SGDP samples varied widely in the number of APG sequences they contained; 4 of the Africans and 4 of the Europeans contained ~1000 APG sequences each, while 5 Africans and 1 European (English) sample contained ~700 insertions (Supplementary Table 5). This could be due to admixture in the CAAPA samples, in which 9 of 19 cohorts were African-American, or to admixture in the European SGDP samples, or some combination of both.

To further examine how well the APG contigs represent continental African populations, we additionally examined the APG contigs present in only the continental African SGDP samples and only the European SGDP samples, but not both. The European SGDP samples cumulatively contained 4,645 of the APG insertions, while the African samples contained 4,381 insertions, with 1,961 of these insertions appearing in both populations. We examined the 2,684 present in the 12 European samples but not the African samples, as well as 2,420 present in the African but not European samples. Although these sequences may not be fully specific to these populations because we examined only 12 samples, we took these to represent sequences tending toward European and African specificity. The European-specific contigs were found at a somewhat lower frequency in the CAAPA samples (47 contigs per individual on average) than the African

specific contigs (63 per individual on average). This difference provides some evidence that despite the admixed nature of the dataset, the APG sequences reported represent sequences present in African populations more so than European populations, though the inclusion of European derived DNA is expected in the APG sequences due to admixture.

Supplementary Note 3

Examination of novel HX1 sequence. We aligned the pan-genome sequences to the 12.8 Mb of novel sequence from HX1 and separately to the entire HX1 genome. We confirmed that the 68.1 Mb of sequences do align to HX1 (with ≥ 90% identity and ≥ 80% coverage for each contig), and these same sequences align poorly or not at all to GRCh38. For example, CAAPA_26854, a 15,617-bp APG sequence, aligned from positions 386–15617 (97.5% coverage) to HX1 Super-Scaffold_142 from positions 1338365–1353606 at 99.7% identity. To verify this sequence was novel, we aligned it to GRCh38.p10 using BLAST [26] and nucmer [34] in addition to bwa-mem [28]. The best match was an alignment of just 425 bp at 81.9% identity, demonstrating that this sequence is essentially absent from GRCh38. However, Super-Scaffold_142:1338365-1353606 was not included in the 12.8 Mb of novel sequence reported as unique to the Chinese genome [14].

Supplementary Note 4

Repetitive regions and linking mates. Several genomic regions have an overabundance of APG contigs tentatively linked to them via mate alignment information, though the linkages were not strong enough to meet our placement criteria (Supplementary Table 3). We examined genomic locations to which >100 contigs linked in five randomly selected samples. All regions in all five individuals had excessively deep coverage, ranging from ~10 times to ~3000 times greater coverage than the 30-40X expected for these datasets. We also observed that all regions had an abundance of mismatches in the alignments. These mismatches in the read pile-ups indicate that these sequences occur in many copies throughout the genome, with somewhat diverged sequence, but are presented by only a few copies in the GRCh38 reference genome. Many of these sequences occur in or near centromeric regions, in which this phenomenon has been previously detected [37]. Since the contigs linking to these regions did not meet our merging criteria, and the locations linked to are not high confidence, the contigs were left as separate sequences in our APG set.

Another consideration in assessing the accuracy of the putative linking mates is repetitive genomic regions, even if the regions are not at deeper than expected coverage, as repetitive sequences are expected, since GRCh38 has over 50% repetitive content. We ran RepeatMasker (with species set to human, using the rmblastn algorithm) on the APG contigs, separately considering the placed and unplaced contigs. As might be expected since non-repetitive sequence is more easily anchored unambiguously, the placed contigs were not as repetitive overall as the unplaced contigs. RepeatMasker masks 61% of the placed sequence as repetitive, half of which (31%) was made up of simple repeats, with most of the remaining repetitive sequence made up

of LINE, SINE, and LTR elements. The unplaced sequence was more repetitive, with an overall masking of 88%, where again the largest category, at 64%, was simple repeats, with another 22% classified as satellites. Mate information linking contigs to regions containing simple repeats or other repeat elements are less reliable links than those anchored in unique sequence, but still serve to provide tentative placements.

Supplementary Note 5

Infection discovery. We examined the classifications of all contigs classified as viral or as *Plasmodium* to determine if any individuals appeared to have viral infections or malaria. All contigs with these classifications from Centrifuge or Kraken were screened for false positives by running blastn to align the contig to NCBI's nonredundant nucleotide database. Only contigs where all top BLAST hits covered at least 95% of the contig at an e-value less than 10^{-20} were considered to be true hits. This resulted in several contaminants of interest, including human betaherpesvirus and malaria. Since only assembled contigs were screened, all contaminants discovered were assembled into at least one contig of a minimum size of 1 kb with some individuals containing hundreds or thousands of assembled contaminant contigs, likely indicating a highly active infection (Supplementary Table 7).

Supplementary Note 6

Primary commands and parameters.

Bowtie 2 alignment, per sample

```
bowtie2-build [GRCh38_no_alt] [GRCh38_no_alt_idx]
bowtie2 -x [GRCh38_no_alt_idx] [reads1] [reads2] > [alignments.bam]
```

Extraction of unaligned reads (and mates) via samtools, per sample
---
```
samtools fastq –f 12 [alignments.bam] -1 [mateUnmapped_R1.fq] -2 [mateUnmapped_R2.fq]
samtools fastq –f 68 –F 8 [alignments.bam] > [R1_mateMapped.fq]
samtools fastq –f 132 –F 8 [alignments.bam] > [R2_mateMapped.fq]
samtools view –f 8 –F 4 [alignments.bam] > [GRCh38Links.bam]
```

MaSuRCA assembly, per sample

```
masurca_config.txt:
**********************************************************
DATA
PE= pe 300 50 [mateUnmapped_R1.fq] [mateUnmapped_R2.fq]
PE= s1 300 50 [R1_mateMapped.fq]
PE= s2 300 50 [R2_mateMapped.fq]
END

PARAMETERS
GRAPH_KMER_SIZE=auto
USE_LINKING_MATES=1
KMER_COUNT_THRESHOLD = 1
NUM_THREADS=24
JF_SIZE=200000000
DO_HOMOPOLYMER_TRIM=0
END
**********************************************************

masurca masurca_config.txt && ./assemble.sh
```

Centrifuge, per sample

```
centrifuge --report-file [centrifuge.report] -x [centrifugedb] -k 1 --host-taxids 9606
-f [masurca_contigs_over1kb.fa] > [centrifuge.output]

centrifuge-kreport -x [centrifugedb] [centrifuge.output] --min-score 0 --min-length 0
> [centrifuge.krakenOut]
```

*** `centrifuge.krakenOut` was used to filter any non-chordate identified reads. ***

RepeatMasker on assembly contigs, per sample

```
RepeatMasker -nolow -species human [filteredContigs.fa]
```

Bowtie 2 alignment of reads to contigs, per sample

```
bowtie2-build [filteredContigs.fa.masked] [contigIdx]
bowtie2 -x [contigIdx] -U [mateUnmapped_R2.fq],[mateUnmapped_R2.fq] -S
[readContigAlignment.sam]
```

Linking mates to implicated region, and aligning to region, per sample

---

```
samtools view -h -F 256 [readContigAlignment.sam] | samtools sort - -n –O bam |
bedtools bamtobed -i stdin | awk '{OFS="\t"}{print $4,$1,$6,$2,$3}' | sort >
[readContigAlignment.txt]

samtools view -H [GRCh38Links.bam] | cat - <(awk 'FNR==NR{main[$1]=$0;next} $1 in main
{print main[$1]}' <(samtools view [GRCh38Links.bam]) [readContigAlignment.txt]) |
samtools sort -n –O bam | bedtools bamtobed -i stdin | awk '{OFS="\t"}{print
$4,$1,$6,$2,$3}' | sed -e 's/\/[1-2]//g' | sort > [matchedMates.txt]

join -j 1 [readContigAlignment.txt] [matchedMates.txt] > [mateLinks.txt]
```

*** Filtering was performed here using python scripts to examine links to contig ends only, and filter based on described unambiguity criteria (see methods). Contig ends and GRCh38 regions meeting criteria were extracted with `samtools faidx` ***

```
nucmer --maxmatch -l 15 -b 1 -c 15 -p [deltaFile] [GRCh38Regions.fa]
[filteredContigEnds.fa]
```

Clustering of placed contigs

```
bedtools merge -d 100 -c 4 -o distinct [placedCtgLocations.bed] > [mergedClusters.bed]
nucmer -p [deltaFile] [repCtg.fa] [restOfClusterCtgs.fa]
nucmer -p [deltaFile] [verifiedClusterCtgs.fa] [unplacedCtgs.fa]
```

Left/Right one end placement merging into two end placement

```
nucmer --maxmatch --nosimplify -p [deltaFile] [leftEndedPlaced.fa] [rightEndPlaced.fa]
show-coords -H -T -l -c -o [deltaFile] > [coordsFile]
```

Removal of redundant placements

```
nucmer --maxmatch --nosimplify -p [deltaFile] [allPlaced.fa] [allPlaced.fa]
```

Clustering of unplaced contigs

```
nucmer --maxmatch --nosimplify -l 31 -c 100 -p [deltaFile] [unPlaced.fa] [unPlaced.fa]
show-coords -H -T -l -c -o [deltaFile] > [coordsFile]
```

*** Additional analysis was performed on the alignments to find and remove contigs contained within two contigs with the ends overlapping (see methods) ***

Further screening

```
kraken --db [database] [APG_Sequences.fa]

blastn -db [nt] -query [kraken_nonMamalHits.fa] -outfmt "6 qseqid sseqid pident length
mismatch gapopen qstart qend sstart send qlen slen evalue bitscore qcovs qcovhsp
staxids sscinames" -max_hsps 1 -max_target_seqs 1 –out [blastOutput]

bwa index [GRCh38.p10_primaryChrs]
bwa index [GRCh38.p10]
bwa mem [GRCh38.p10_primaryChrs] [APG_Sequences_noContamiants.fa]
bwa mem [GRCh38.p10] [APG_Sequences_noContamiants.fa]
```

Genotyping, per sample

```
bwa index [APG_Sequences_final.fa]
bwa mem [APG_Sequences_final.fa]  [contigsFromMaSuRCA.fa] > [sampleToAPGAlignment.sam]
```

Comparisons to other genomes

```
bwa index reference
bwa mem [reference] [APG_Sequences_final.fa]
```

Reference genomes used

```
 GRCh38_no_alt  GCA_000001405.15_GRCh38_no_alt_analysis_set.fna
    GRCh38.p10  GCF_000001405.36_GRCh38.p10_genomic.fna
         KOREF  GCA_001712695.1_KOREF1.0_genomic.fna
           HX1  hx1f4s4full_3rdfixedv2.fa
```

In addition to these primary commands, additional filtering steps and custom analyses were performed, as described in the methods section. Filtering commands were primarily performed using `awk` and filtering for identity and coverage was always performed on `coords` files produced by `show-coords`; if `bwa` was used for alignments in place of `nucmer`, the `sam` files were converted to `nucmer` delta files, using the CIGAR strings and lengths to determine identity and coverage.

Supplementary Tables 5-7

Supplementary Table 5 | APG contig presence in Simons Genome Diversity Project individuals

Sample ID	Population	Country	Sex	Number of APG Contigs Present
LP6005442-DNA_E10	English	England	M	796
LP6005442-DNA_F10	English	England	F	680
LP6005441-DNA_A05	French	France	M	963
LP6005441-DNA_B05	French	France	F	810
LP6005441-DNA_C11	Sardinian	Italy	M	943
LP6005441-DNA_D11	Sardinian	Italy	F	905
LP6005442-DNA_A11	Spanish	Spain	M	817
LP6005442-DNA_B11	Spanish	Spain	F	1011
LP6005442-DNA_C10	Finnish	Finland	M	893
LP6005442-DNA_D10	Finnish	Finland	F	892
LP6005442-DNA_A08	Hungarian	Hungary	M	1041
LP6005442-DNA_B08	Hungarian	Hungary	F	1007
LP6005441-DNA_G08	Mozabite	Algeria	M	1034
LP6005441-DNA_H08	Mozabite	Algeria	F	980
LP6005443-DNA_A01	Bantu	Kenya	M	791
LP6005441-DNA_B02	Bantu	Kenya	F	991
LP6005442-DNA_G10	Gambian	Gambia	M	710
LP6005442-DNA_H10	Gambian	Gambia	F	690
LP6005442-DNA_G11	Mende	Sierra Leone	M	720
LP6005442-DNA_H11	Mende	Sierra Leone	F	711
LP6005592-DNA_C03	Mbuti	Congo	M	690
LP6005441-DNA_B08	Mbuti	Congo	F	914
LP6005442-DNA_A02	Yoruba	Nigeria	M	925
LP6005442-DNA_B02	Yoruba	Nigeria	F	980

Twenty-four individuals from the Simons Genome Diversity Project from 12 populations, 6 African and 6 European, were examined to determine presence/absence of the APG contigs. Each individual's assembled contigs were aligned to the APG contigs to determine the number of APG contigs present in the individual.

Supplementary Table 6 | Cohorts of CAAPA samples.

Cohort	Number of Samples
African American (Atlanta)	50
African American (Baltimore-DC)	50
African American (Chicago)	50
African American (Detroit)	50
African American (Jackson, MS)	50
African American (Nashville)	48
African American (NYC)	48
African American (San Francisco)	50
African American (Winston-Salem)	50
Barbados	49
Brazil	47
Colombia	50
Dominican Republic	47
Gabon	34
Honduras	50
Jamaica	50
Palenque	34
Nigeria	50
Puerto Rico	53

Data was collected from 19 distinct cohorts across the Americas, the Caribbean, and Africa resulting in 910 analyzed samples.

Supplementary Table 7 | Contigs assembled from contaminants of interest.

Sample ID	Population	*Plasmodium falciparum* (# contigs)	*Plasmodium malariae* (# contigs)	*Human betaherpesvirus 6B* (# contigs)
LP6005271-DNA_C04	Gabon	4184	-	-
LP6005271-DNA_D04	Gabon	3	-	-
LP6005271-DNA_E04	Gabon	2615	-	-
LP6005271-DNA_A03	Gabon	2	-	-
LP6005271-DNA_A04	Gabon	1	-	-
LP6005271-DNA_B03	Gabon	4077	-	-
LP6005271-DNA_C02	Gabon	6	2	-
LP6005271-DNA_C03	Gabon	2	-	-
LP6005271-DNA_F02	Gabon	36	-	-
LP6005092-DNA_B02	Nigeria	3	-	-
LP6005092-DNA_E03	Nigeria	8	-	-
LP6005092-DNA_H02	Nigeria	1	-	-
LP6005092-DNA_C02	Nigeria	1	-	-
LP6005092-DNA_A04	Nigeria	2	-	-
LP6005092-DNA_C01	Nigeria	1	-	-
LP6005092-DNA_G02	Nigeria	676	-	-
LP6005092-DNA_G03	Nigeria	89	-	-
LP6005092-DNA_H03	Nigeria	2	-	-
LP6005092-DNA_A06	Nigeria	1	-	-
LP6005092-DNA_B05	Nigeria	16	-	-
LP6005092-DNA_F06	Nigeria	15	-	-
LP6005092-DNA_A07	Nigeria	2	-	-
LP6005092-DNA_B06	Nigeria	7	-	-
LP6005092-DNA_B07	Nigeria	-	10	-
LP6005092-DNA_D06	Nigeria	5	-	-
LP6005092-DNA_E02	Nigeria	6	-	-
LP6005092-DNA_F05	Nigeria	1	-	-
LP6005092-DNA_G05	Nigeria	1	-	-
LP6005092-DNA_H06	Nigeria	1	-	-
LP6005107-DNA_F03	African American (Winston-Salem)	-	-	3

Contigs from *Plasmodium falciparum*, *Plasmodium malariae*, or *human betherpesvirus 6B* were assembled in 30 individuals. Though most *Plasmodium* contigs detected were falciparum, one individual had contigs present from both *Plasmodium* species and one solely from *malariae*. All individuals with *Plasmodium* contigs were from either the Gabon or Nigeria cohorts.

Supplementary Methods

Exact Matching of Contig Ends

One a region was unambiguously determined for a contig end, we performed NUCmer alignments to determine the exact placement location. For unambiguous contig ends, we aligned the terminal 200 bp of sequence to the region determined by the mate placements. These regions were up to 2 Kb in length, which we padded with an additional 500 bp taken from both sides of the region. Alignments were performed without repeat-masking because the region had already been unambiguously identified. We used the parameters `--minmatch 15 --breaklen 1` to disallow gaps or mismatches in the alignments and left all other parameters as defaults. If we found at least one exact match of at least 15 base pairs within 5 bases of the contig end, and all exact matches were consistent with one another, an exact breakpoint was determined by chaining the alignments. The resultant aligned portion of the contig was recorded (to be trimmed off later in the pipeline) and the endpoint of the alignment was recorded as the insertion location for that end of the contig.

Incorporation of PopIns Output

PopIns was run on MaSuRCA assemblies beginning with the `popins merge` step, through the `popins place-finish` step. In the `popins merge` step, PopIns produces new contigs by merging those provided to it into new merged contigs. To obtain clusters of placed contigs which could be more easily merged with our pipeline, we aligned all MaSuRCA contigs to each merged contig created and placed by PopIns. We grouped contigs that fully aligned with over 98% identity into a single cluster representing one insertion and location. For all contigs in each

cluster, we then aligned each contig's ends to the placement location with NUCmer, as described above, to attempt to determine its exact placement.

Notably, PopIns generates placements of a single end of a contig using the VCFv4.2 breakend format to specify how much of a merged contig is inserted at a breakpoint. Thus in many cases PopIns output several placements for the same contig that did not agree in orientation or placement location. If we could not verify a placement made by PopIns via our independent alignment of the contig ends to the placement region, we excluded it from the final set of insertions. Of the contigs in a cluster that could be exactly placed with NUCmer, if one or multiple contigs had both ends placed, the longest of these was reported as the representative of the cluster and the insertion was added to the set of two-end placed insertions.

If no single contig had both ends placed by NUCmer, up to two representatives were chosen, with the longest contig being chosen as the representative for each end of the contig. This resulted in the potential for two separate one-ended clusters, which were then added to the one-end placed insertion set as follows. All PopIns representatives where the contig had already been placed in a two-ended placement cluster were excluded regardless of location conflicts, as the two-ended clusters necessarily had more evidence supporting them. We further excluded as redundant any PopIns placements within 100 bases of an existing one-ended or two-ended placement location.

Once PopIns placements were incorporated into the one and two end placement sets and clusters had been finalized (see Methods), we attempted to verify contig placements produced by PopIns.

To verify the placements we examined the placement locations of linking mates from all contigs in the cluster of a PopIns placement. Clusters in which fewer than 25% of all linking mates aligned within 5 Kb of the GRCh38 placement location were removed. If no mates existed, the cluster was not removed. This resulted in the removal of a number of PopIns placements, including several for which we had determined a one-ended placement with very strong mate-pair support but a PopIns' two-ended placement disagreed.

One End Placement Merging

We ran `nucmer --maxmatch --nosimplify`, followed by `show-coords -o` (with annotation) to align representative contigs of clusters within 500 bp that were candidates for merging. If NUCmer annotated the representatives as identical, or if it found that either contig contained the other with at least 97% identity, we merged the clusters and reported the longer representative contig. In cases where the ends of two contigs overlapped in the correct arrangement and orientation relative to their placements, we merged the overlapping ends by extending the sequence of the longer contig with that of the shorter contig as indicated by the alignments. The resulting merged sequence and cluster was then moved to the 2EP set. If NUCmer identified other alignments between representatives covering at least 50% of one of the representatives and the clusters shared any contigs (i.e. at least one contig was contained in both representatives), the clusters were merged. However, because these representatives were more divergent the representatives were not merged and the longer representative was reported in the 1EP set (Supplementary Table 1).

Alignment of Final Sequences to GRCh38.p10

Although the reads used to assemble these contigs had initially failed to align to the genome, in some cases the resulting contigs had sufficient similarity that they could be aligned. This resulted in the removal of five two-ended placements, 24 one-ended placements, and 249 unplaced contigs. Among the 29 placed contigs that were removed at this step, all had alignments between 90% and 93% identity and were present in 10 or fewer samples; this may indicate that some individuals simply had slightly more divergent sequence than the overall population, explaining why Bowtie2 failed to align their reads initially. The best alignment locations that had at least 50% of the contig aligned to GRCh38.p10 at ≥ 80% identity were determined by taking alignments to both primary, alternate, and patch sequences, and calculating a score by multiplying the percent identity by the alignment length and are reported in Supplementary Tables 1-2.

Supplement Only References

37 Miga, K. H., Eisenhart, C. & Kent, W. J. Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments. *Nucleic Acids Res* **43**, e133, doi:10.1093/nar/gkv671 (2015).